CERTIFICATION OF HIGHLY AUTOMATED VEHICLES FOR USE ON PUBLIC ROADS

John McDermid
University of York and FiveAI
UK

Phil Koopman
Carnegie-Mellon University
USA

Robert Hierons
Brunel University
UK

Siddartha Khastgir
University of Warwick
UK

John A Clark
University of Sheffield
UK

Michael Fisher
University of Liverpool
UK

Rob Alexander
University of York
UK

Kerstin Eder
University of Bristol
UK

Pete Thomas
University of Loughborough
UK

Geoff Barrett
UK

Philip Torr
University of Oxford and FiveAI
UK

Andrew Blake
FiveAI
UK

Subramanian Ramamoorthy
University of Edinburgh and FiveAI
UK

Paper Number 19-0343

ABSTRACT

Objective: A number of different methods must be combined for the robust certification of highly automated vehicles (HAVs) for deployment in ODDs encompassing public roads. This paper, which is authored by a braintrust of the world's leading academics in validation, verification and certification and affiliated with Europe's largest autonomous vehicle developer FiveAI, proposes a core set of processes.

Methods: The paper discusses in detail: (1) requirements discovery; (2) behaviour requirements; (3) simulation as a tool for verification; (4) useful tools and methods.

Results: We propose a process centred around hyper-scale fuzzed scenario-based testing and the use of coverage driven verification methods in digital twins of the ODD and using generative models representative of each ODD. Testing must cover both full stack testing, which will require photo-realistic and sensor-realistic rendering of scenarios and objects, together with accurate sensor modelling and motion planning stack testing, will require robust beliefs over scenario actor behaviours to test predictive, planning and motion synthesis.

Discussion and Conclusions: The paper poses several questions for policy makers: (1) Could a validation, verification and certification system that incentivizes sharing of scenarios while protecting the value intrinsic to their discovery, improve safety across the industry? Could it be used by an approval body such as a national Certification Agency to establish a high standard for national certification? (2) Can the industry agree on a scenario description language that supports coverage-driven verification and is extensible? (3) What should the specification of an appropriate simulation environment be? (4) Could the specification for a test oracle be made available and could this be based on a formal description of 'good driving'? (5) Is auditable adherence to the IATF16949:2016 quality assurance process sufficient to satisfy 'Conformity of Production'?

Key questions also remain, including: (a) What machine learning methods should be applied to directed random testing in coverage driven verification? (b) Given the high dimensionality of the test space, what coverage measures are meaningful in generative and ODD digital twin verification? (c) Which computer vision methods can we apply to the 3D reconstruction of digital twin worlds from photogrammetry, LIDAR scans and other modalities that mean accurate, up-to-date digital twins are feasible? (d) What hardware acceleration beyond GPUs can we design and apply to enable faster-than-real-time full stack verification of HAVs? (e) How can we apply formal software checking to the complex integrated systems required for autonomous driving to ensure that each build achieves its goals without bugs or gaps? (f) How do we really apply formal mathematical methods to express the Digital Highway Code (DHC), vehicle dynamics and other road user expectations and behaviours to verify the behavioural safety of HAVs? (g) How can we verify HAV systems that comprise of one or more end-to-end neural networks with the requirements to explain failure modes and take corrective actions to improve their performance using human readability and intermediate outputs of modular processes? (h) How might we extrapolate randomized testing, including near collisions, into a measure of probability of collision generally?

INTRODUCTION

The development of Highly Automated Vehicles (HAVs) capable of performing SAE Level 4 autonomous driving (AD) is a huge attractor of intellectual and capital investment across the world today. No single company has developed a complete system of hardware and software that can realistically be deployed in unconstrained urban environments yet, but several teams expect to attain a standard that they consider should be deployable in our cities within 3-5 years, perhaps sooner in simple, wide, well-lit and sparse urban zones.

If we are able to deploy such systems, they have the potential to unlock major economic and societal benefits for city dwellers and our economies as whole: the means to offer low cost universal demand-responsive mobility to every citizen, a tool for unlocking unproductive commuting time and enabling economic engagement for everyone, increased road safety for all road users and materially lower pollution, congestion and resources today wasted in manufacturing, assembling, parking and disposing of personal vehicles. In delivering safe, on-demand, shared end-to-end journeys across a city, including zero occupancy dispatch, the advent of HAVs will presage an inevitable and

dramatic shift away from the car ownership model which has prevailed for over 100 years towards a low cost, shared mobility-as-a-service world which can integrate more successfully with shared public transport services.

Reaching that goal is requiring large teams of computer scientists, mathematicians, engineers, roboticists and manufacturers to work together across disciplines to meet many engineering and implementation challenges. These are highly complex technologies which must be built using commercially feasible hardware and by selecting, hardening and integrating many recent, and still being developed, cutting-edge research breakthroughs from academia. Not only must those teams themselves test and validate the resulting systems to ensure they do not cause injury or death to humans and animals or endanger property, the public as a whole – along with their elected representatives – need the reassurance of a clear and rational means to validate the safety of these systems independently through an appropriate regime of validation and a process of certification.

What is already clear is that the combinatorial effects of different road topologies, road users, appearances, lighting, weather, behaviours, sensors, seasons, velocities, randomness and deliberate actions cannot be adequately experienced in on-road testing alone, even in the constrained operational design domain (ODD)[1] constraints implied by level 4 autonomy. And if they could, it would take billions of highly varied miles of drive testing on each individual build of hardware and software to reach a statistically reliable level of validation that the proposed system could even attain human levels of safety. The stakes are high: even a single bad character in one line of software code can, and has, caused catastrophic failure. Any such failures found in testing would require build change with the potential for regressive effects meaning that the testing process would need to start again. This means that a methodology that rests solely, or mainly, on on-road testing is infeasible.

Although international and national standards exist for the functional safety of individual components and sub-systems (e.g. ISO 26262), no regulatory authority today has a well-defined system for validation of HAVs as a complete system rooted in an understanding of the problem space. For example, measures set out by California's Department for Motor Vehicles (DMV) include the reporting of 'driven miles per disengagement' and this is sometimes presumed as a competitive measure of maturity and safety of proposed systems. Not only can these measures be statistically meaningless[2] – only a tiny fraction of the potential state space has been explored – but they are potentially harmful in encouraging premature and non-representative on-road testing, discouraging interventions and propagating a misleading perspective on safety, leading to loss of life. Our opportunity is to set out a framework and ensure our testing and validation processes are world-leading, thereby ensuring safety for our citizens, gaining economic advantage first and unlocking global business opportunity for our scientific and engineering companies who embrace the regime.

In light of the perceived benefits from HAVs, the UK government has indicated a willingness to provide an exemption from (or modification to) the construction and use regulations so that they can be used on public roads before 2021, with more wide-sweeping legislative reform targeted thereafter.

While nobody should expect these systems to be perfect, we should expect them to reach human driver safety levels and to be progressively tightened to significantly higher safety levels over time. In that quest, it's important that development and testing practices are established, are followed and that developers and regulators can measure the safety performance of each combination of technologies in representative environments to a statistically meaningful standard.

This paper therefore seeks to explore the problem space, propose appropriate practices and contribute to the establishment of a certification regime that will safely unlock the value of HAVs to our citizens.

SAFETY OBJECTIVE

Once human levels of safety have been attained and surpassed, a primary objective for any HAV program should be to reduce the incidence of injury to humans, animals and property.

The UK's Department for Transport reported that in 2016, 327 billion vehicle miles were driven in the UK and there were 137,000 'accidents' reported to the police which are essentially collisions reported to insurance companies. That equates to one reported 'accident' every 2.4 million driven miles, with reported serious injury occurring every

12.6 million driven miles. Not all incidents are police-reported 'accidents' and collisions are likely to be significantly more frequent than the reported statistics indicate. The number of motor insurance claims in 2016 was 4.34 million[3], or one claim every 75,000 driven miles and many accidents are not reported to insurers.

The Institute of Advanced Motorists in their 'Licensed to Skill' report in 2010 estimated that around 94% of these incidents can be traced directly to human error but, in terms of public acceptance, HAVs that cause human injury or death are likely to be held to a higher standard than fellow humans, however irrational that may be. We do not know exactly how much higher these standards will need to be, but it seems likely that in order for the general public to accept the relinquishing of control to these emerging autonomous systems, halving the collision rate would be a minimum target. And that implies an incident rate of around once per 200,000 miles.

If the above were achieved, HAVs could halve serious injury and death on our roads, saving over a thousand lives each year in the UK alone, most in the 15-29 age bracket. The societal benefit is an overwhelming one.

REGULATORY CONTEXT

Regulations governing the standards, testing and certification of product conformity of vehicles on public roads in Europe are governed by European Union (EU) Directives and, by virtue of the fact that the EU is a contracting party to global technical regulations coordinated by the United Nations (UN), are also governed by safety and environmental aspects of UN regulations too. The UN regulations are managed by the World Forum for Harmonization of Vehicle Regulations, a permanent working party of the United Nations Economic Commission for Europe (UNECE). UNECE and EU countries take part in the technical preparatory work of the Forum and UNECE exercises the right to vote in the Forum on behalf of the EU.

Directive 2007/46/EC provides that EU countries share a common legal framework and general technical requirements for the approval of new vehicles and of systems, components and technical units designed for them. It establishes a harmonized framework so as to facilitate the registration, sale and entry into service of new vehicles anywhere in the EU, as well as rules regarding the sale and entry into service of vehicle parts and equipment.

For vehicles to be approved for registration, sale and entry into service, the 'whole vehicle' must pass all applicable approvals and, for this purpose, a single production sample is selected and tested as representative of the type to be approved, hence the term Type Approval. In order to gain whole vehicle Type Approval, each of the various systems, e.g. brakes, emissions, noise, etc., must be tested and meet the standards set out in the relevant EU Directives and UNECE regulations. There are no additional whole vehicle tests; instead the sample vehicle will be considered as a whole by a designated approval body and if the production sample of the complete vehicle can be confirmed to match the specifications contained in all the separate Directive approvals, then on submission of the relevant manufacturer's information documents, it will result in the issue of a European Whole Vehicle Type Approval Certificate (EWVTA).

EU Regulations permit any EU Member State to appoint an Approval Authority to issue those EWVTAs and to appoint a Technical Service to carry out the testing to the EU Directives and Regulations standards. In the UK, both the Approval Authority and Technical Service functions are performed by the Vehicle Certification Agency (VCA).

No technical Directive yet exists for the approval of HAVs. Moreover, existing Directives sometimes conflict with such operation: one example being the UNECE regulation no 79 on steering type approval which places an effective 12km/h limit on HAVs through clause 5.1.6.1 which states that 'Automatically Commanded Steering …. action shall be automatically disabled if the vehicle speed exceeds the set limit of 10 km/h by more than 20 per cent'.

Several working groups have been established to seek consensus on how UNECE and EU Directives should be amended to permit HAV operation.

OBJECTIVE OF THIS PAPER

The objective of this paper is not to identify conflicts within the existing Type Approval process or suggest amendments to the existing Directives – this work is already underway through the various working groups – but to

identify the key components of a validation, verification and certification process for HAVs that could be adopted to ensure their safe introduction on UK and European roads, along with highlighting the open research questions in relation to that process.

Until now, there have been no universally accepted dividing lines between validation (which checks that the required specification is complete and accurate), verification (which is the process used to gain confidence in the correctness of a design or system with respect to its specification) and certification (which is the legal recognition by a certification authority that a product or service complies with the requirements).

This paper therefore proposes the following approach:

- To propose a target general framework, to be achieved over time, which is capable of being applied to the discovery and establishment of adequate specifications for HAVs, which defines a process for validating those specifications (including safety properties) and which establishes a means of verifying that any candidate System under Test (SUT) is robust to all major classes of defects against those validated specifications to a measurable standard. This will permit HAV technology developers to attain and, over time, exceed human levels of driver safety

- To establish that framework as a code of practice that the UK (and by extension, if adopted, the EU) will require HAV technology developers to internally adopt for V&V, if they want to deploy in those regulatory environments

- To require that the UK (and by extension, if adopted, the EU) certification authorities adopt the same framework to independently verify a randomized subset of the design verification

- To require certification authorities (or an approval body) to conduct an audit of the quality assurance (QA) processes of the HAV technology vendor to ensure that the design and test methodology they employ is rigorous

- To require that there is a process whereby HAV technology developer software updates remain robust to regressions and conform to specifications as they are updated

This paper is structured in the above order, first identifying the key attributes of a V&V framework for HAVs and then discussing how this may be applied in the context of certification.

A FRAMEWORK FOR HIGHLY AUTOMATED VEHICLE SAFETY VALIDATION & VERIFICATION

To ensure a HAV validation framework can establish and measure performance against necessary safety standards, it must address at least five types of defects. These are intended to cover all types of potential faults in the system, its environment or its use:

- **Requirements defect**: the system is specified to do the wrong thing (defect) or is not required to do the right thing (gap) or the Operational Design Domain (ODD) description is incomplete (gap) or inaccurate (i.e. a validation defect). These types of defects may manifest as product defects where the system does something unsafe or as process defects, i.e. where there is insufficient evidence of safety

- **Design defect**: the system design fails to meet a specified safety and/or functional requirement or fails to respond properly to violations of the defined ODD

- **Implementation defect**: the implementation of the system does not conform to its design specification

- **Verification plan defect**: the verification plan fails to exercise potential states (e.g. corner cases) in requirements or to identify instances in which the vehicle's interpretation of the external world is incorrect to the degree that safety is impaired

- **Safety or reliability defect**: an invalid input or a corrupted system state causes an unsafe system behaviour or failure (e.g. sensor noise, component fault, software defect) or an excursion beyond the ODD due to external forces

HAV Requirements
A key challenge for the safety assurance of HAVs is in understanding the system requirements and validating that they sufficiently represent the ODD before verification of the system against those requirements can begin.

Vehicle Road Testing for Requirements Discovery Discovering the system requirements for HAVs in a target ODD is a huge and necessarily incomplete task, partly because the real world has high dimensionality and combination possibilities – objects, environment, behaviours, degradations, sensors, occlusions and so on – but also because the process of discovering precisely what is needed is never finished as the real world keeps changing.

Since no digital record exists anywhere that does or could possibly describe all the possible stand-alone and combinatorial possibilities that might exist in anything other than the simplest ODD the HAV could be presented with, any system specification will inevitably still present gaps to the real requirements.

Minimizing those requirement gaps is the primary motivation for on-road vehicle data-gathering and testing operations. These include:

- Detecting novel road hazards
- Detecting lighting, weather, specularities, sensor combinatorial failures in the ODD
- Discovering behaviours that violate normal traffic rules and finding exceptional but possible scenarios
- Learning accepted norms of driving
- Discovering unusual road user configurations, surfaces, aesthetics and behaviours
- Discovering how behaviours vary by time of day, weather, season
- Finding situations where sensing modalities fail, localization exhibits randomness or biases
- Finding and correcting misleading but well-formed map data
- Discovering types of novel road signs and traffic management mechanisms specific to a micro-location or event
- Finding unusual road markings and vandalism, degradations, mistakes
- Learning emergent traffic effects caused by the HAV and learning third-party behaviours due to the presence of the HAV
- Learning malicious third-party behaviours

Once a discovered requirement is identified by vehicle testing in the ODD and validated (distinct from an SUT verification failure against an existing system specification), there should be an update to the system requirements for that ODD, an update to the requirements for the fidelity of the simulation environment, the generation of one or more new test cases or a combination of all three.

The larger the ODD, the longer and more expensive the requirements discovery process will be. It is for this reason, amongst others, that we are a long way away from a true SAE level 5 autonomous driving capability.

Hazard Analysis While real-world discovery of requirements is an essential part of requirements capture, systems engineering methods like STPA (Systems Theoretic Process Analysis developed by MIT) or Functional Hazard Analysis (FHA) should also be adopted to better understand where defects of any kind can lead to hazards. STPA has been used in the aviation industry by Boeing, Embraer and NASA.

Encoding Scenario Requirements Efforts are underway in various countries to document the HAV's requirements as a curated set of vehicle behaviours and scenarios, the largest being the **P**roject for the **E**stablishment of **G**enerally **A**ccepted quality criteria, tools and methods as well as **S**cenarios and **S**ituations for the release of highly-automated driving functions supported by Germany's Federal Ministry for Economic Affairs and Energy (the **Pegasus** project).[4]

The capture and curation of such scenarios and behaviours provides a means not just to specify system performance but to develop and verify functionality that attempts to meet those system specifications. Such scenarios and behaviours can also be used to generate regression tests which can be replayed in simulated worlds to play a part in verifying some aspects of the behaviour of an entire system-under-test (SUT), as well as to provide a baseline from which to randomize variables to discover new failure modes of the SUT.

The Pegasus project, which has gained the broad support of many major participants in the German automotive industry, has the aim of developing procedures for the testing of AD functions, in order to facilitate the rapid implementation of HAVs into practice.

Scenarios are a key element of the Pegasus verification concept in that they are the basis for eliciting whether the HAV under test exhibits appropriately safe system-level behaviours. Scenarios have a functional view (described in free text), a logical view (with a set of ranges for the "interesting variables"), and a concrete view (with all these variables given concrete values).

Pegasus scenarios can be captured in a number of different ways but since the whole project is still at an early stage, today there is just one live capture method, **OpenSCENARIO**. This is an XML-based format proposed by Vires Simulations Technologie GmbH and capable of being interpreted on Virtual Test Drive, a widely-used simulation platform Vires has developed and marketed. OpenSCENARIO is therefore currently being adopted by all participants in Pegasus as a pro tem standard for capturing concrete test cases. Longer-term, the Pegasus project hopes that OpenSCENARIO might evolve to become a cross-platform industry-wide standard for encoding scenarios and behaviours that could be ported to many or all simulation and testing execution platforms (including software-in-the-loop simulators, hardware-in-the-loop simulators and test track setups).

Behaviour Requirements

Encoding such scenarios and confirming the ability of the SUT to perform a manoeuvre that avoids collision in testing against each scenario has limitations unless we can also verify whether the SUT can conform to traffic laws and to driving codes of practice during that testing.

Encoding Traffic Law & Driving Behaviours For that, we need a publicly-available, machine-readable and complete set of those traffic laws and driving codes and conventions, a Digital Highway Code (DHC). That DHC must include exception handling rules, for example: when and how exactly can a vehicle cross a centre dividing line, if present, to avoid a lane obstruction; when would it be acceptable to mount a sidewalk; what should a driver be permitted to do if traffic lights are defective and so on. These conventions should extend to polite behaviour on the road in that jurisdiction, including when a HAV should let other road users merge into its lane, to what extent does the HAV have a responsibility to ensure the most efficient use of the road network, etc.

Simulation as a Tool for Verification

The core of any effective verification program for HAVs will be the use of simulation.

One or more simulators must be developed to be capable of replaying scenarios in which the road, lighting, weather, degradations, objects, road user actions and interactions can be re-generated and used for verifying the SUT. How complete and representative of the real world a simulator needs to be depends on which parts of the SUT stack is being exercised and tested and how much testing is necessary to explore the state space. But at some level, the full SUT stack must be tested, which means that photo-realism, radar and Lidar reflectivity, sensor models, vehicle dynamics, road surface, human actions must be model variables on top of a baseline simulation capability. A critical issue is how good these simulations must be in order to be effective verification tools, when considered in conjunction with other verification methods including lab test, real-world driving, etc.

Components An example of the principal components of a simulation model suitable for HAV verification is shown in figure 1.

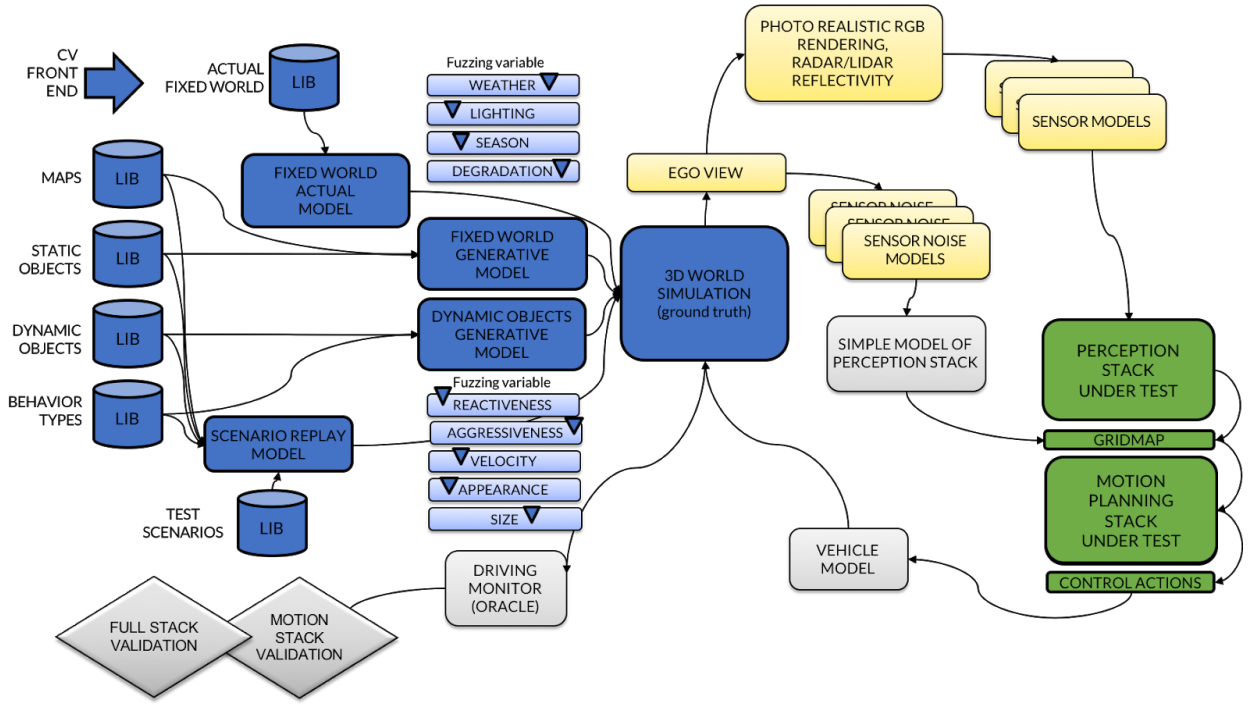

Figure 1: Example simulation model to support HAV verification

Scenario Replay The base point for using a simulator is the interpretation of a scenario and its re-creation in a simulated world, including the instantiation of all recorded dynamic agents and behaviours captured in that scenario. Given a model of road, layout, obstacles, occlusions, road users and behaviours, a simulator can test the predictive capability of the SUT stack, test its ability to plan the HAV motion in the context of road rules and uncertainties, set a trajectory and control the HAV safely, given a model of the HAV vehicle dynamics. A noise model is used to inject perception stack uncertainty into measurements and, at a base level, testing would confirm safe operation given that scenario and noise. Such behavioural tests could run much faster than real-time since the perception layers of the stack would be replaced with noise models. Types of noise as well as scene and user behaviours are varied to discover new failure modes in a process sometimes referred to as 'fuzzing'.

Coverage Driven Verification In addition, simulation models can also be generative in the construction of new scenarios that are theoretically possible but have not yet been captured in road testing or in manual test case generation. Generative models allow the exploration of state space beyond fuzzing, either on a random or directed basis. Several useful techniques exist for exploring and finding new failure modes. These include (i) coverage metrics which then can direct random test generation to broaden that coverage and (ii) using machine learning to reward finding new combinations of layouts, objects, behaviours, velocities, lighting, weather, season etc. that cause the SUT to fail using the proximity of close or actual test failures from random test generation. These types of verification are usually referred to a coverage driven verification. Since the state space being explored for HAVs is extremely large, to all intents and purposes infinite, coverage driven verification using generative modelling would emphasize exploration over density of test coverage which would, by definition, remain sparse.

Randomization & Direction Hand-written tests can be created by focusing on expected corner cases and then automatically 'fuzzing' around them. This involves human generation of very general abstract scenarios which are then instantiated into many different concrete scenarios and coverage is typically clustered around these corner cases.

Another approach, currently favored by the Pegasus project, is to use randomization with expected distributions, often referred to as Monte Carlo simulation. This approach provides sparse coverage but will deliver estimates of the frequency of failures.

But directed or machine-learning randomization is the favored method for defect-finding. These techniques emphasize the capability to reach edge cases, at the expense of overall failure rate estimation and can also be used to support the systematic discovery of defects with respect to adversarial perturbations.[5]

It is important to note that the statistical distribution of test cases has to be driven both by the probability of occurrence and the magnitude of a potential loss (i.e., by risk, not just occurrence). Otherwise a relatively rare scenario that could result in a fatality will be under-represented. In other words, while some testing should concentrate on normal functionality, a substantial portion of testing will need to emphasize infrequent but dangerous situations.

Verification in Digital Twins As well as fuzzed scenario replay and the use of coverage driven verification in generative models, simulation models can also be built that replicate the real world ODD and the likely object types and behaviours in it. Applying coverage driven verification to such digital twins will find additional failure modes which may not be found in fuzzed test scenarios nor easily found in fully-generative models. It can also provide a higher level of coverage that, to some extent, is measurable in relation to the real-world twin the digital model represents. This approach is therefore an extremely useful addition to the first two techniques but requires the creation of a 'digital twin': literally a realistic model of the ODD along with a distribution of dynamic agents and behaviours that are representative, in absolute terms, of that specific ODD.

Full Stack as well as Motion Planning Not only must the predictive, behavioural and control aspects of the HAV stack be tested in simulation (motion planning) but the full stack must be tested too, since many failure modes are possible in the sensors, localization, perception, interpretation, classification and confidence measurement layers of the stack alone, or as they interact with the motion planning layers as a whole.

Full stack testing is a bigger task than behavioural testing, since the simulation model must now render the scene with photorealistic textures, lighting effects, reflections, specularities, shadows, weather and seasons. The same is true for all road users, pedestrians, cyclists and any other objects in the scene, and a library of such objects must be curated and maintained. Sensor outputs need to be modelled based on the placement of the sensors on the HAV and on accurate models of the behaviour of those sensors, including any dynamic range limitations, calibration limitations or errors, quantization, timing delays, race effects, color and lighting sensitivities, blur and other optical perturbations. And it's not just RGB that needs rendering, it's also radar, Lidar and other sensing outputs, given material, density, weather and other conditions. A useful contribution here may also come from the Pegasus project, in the shape of the proposed standard for defining weather, signs, sensor inputs and so on, called Open Simulator Interface (OSI). OSI could be used in the future to connect various simulated artefacts produced by different companies.

Not surprisingly, full stack testing is computationally expensive and may run well below real-time, meaning that attaining meaningful levels of coverage will require substantial datacenter resources to spin up multiple instances of the simulator, sensors and the SUT.

Useful Tools and Methods

A Scenario Language A well-defined language to describe scenarios that can be interpreted by each simulator to re-create essentially the same scenario and behaviours with high fidelity is likely needed. Perhaps OpenSCENARIO is such a language, but if is not suitable a new one will need to be created. One such initiative is being pioneered by Foretellix, an Israeli startup, and they simply call this language Scenario Description Language or SDL. But whether it is OpenSCENARIO, SDL or something different, a cross-industry agreement must be reached.

A Scenario Sharing Library A library of scenario test cases must be developed by or made available to each company building HAV technology. Of course, it needs to be as comprehensive as possible in many dimensions.

A suitable scenario library should:

- Cover cases where collisions are possible and must prove SUT avoidance within the boundaries of the DHC
- Monitor and report behaviour before, during and after collisions or near-collisions
- Pay specific attention to interactions between HAVs and humans and grade for collisions, near collisions, breaches of the DHC and any other behaviours where the SUT adversely affected any virtual passengers, traffic flow or other road users
- Emphasize the verification of inference or deep neural network-based algorithms and find failure cases where interpretability is poor
- Use (i) replayed scenarios (ii) fuzzed replayed scenarios (iii) generative models with directed or machine-learning randomization and (iv) replayed, fuzzed and generative behaviours in a digital twin of representative (or whole) digital twin instantiations of target ODDs
- Look for both “expected” and “unexpected” defects
- Employ configuration files that permit portability from city-to-city through tractable modification of vehicle, sensors, weather files, location, signage, human behaviours, road markings and DHC

These objectives must be met whilst maintaining transparency and maintainability. As the number of test scenarios grows and as they become ever more intricate, this will become a real problem unless tackled from the outset by the industry in initiatives, like the Pegasus project or by the adoption by several key players of a succinct means of encoding those scenarios.

A Motion Language A motion language with qualitative actions (e.g. "follow at a safe distance" or "pull in safely") could add significant value to the validation and verification processes. In the UK, learner drivers are tested on their knowledge of the Highway Code and advanced motorists are encouraged to follow the guidance set out in Roadcraft: The Police Driver’s Handbook.

Codifying what constitutes good driving as described in these manuals, as in the suggested DHC can serve the following uses:

- Engage with the public on target HAV behaviours on UK roads
- Complement scenario-based, coverage driven system testing
- Provide the basis behaviour for a test oracle
- Provide the basis functionality for low complexity monitoring systems to be used by HAVs in run time to improve safety robustness.[6]

Model-checking using a formal method might be used to automatically verify safety properties of the DHC using observation of behaviours from the real world.

Application of Formal Methods Important aspects of hardware design are already amenable to automated proof methods, making formal verification possible to introduce. But the application of these methods to software design is a more complex problem, although a number of mathematical methods do exist for proving a computer program satisfies a formal specification of its behaviour.

As development moves towards higher levels of autonomy then the need for stronger, formal software verification becomes acute. One of the fundamental steps that needs to be taken to understand and analyse HAVs is that we must assess not just what a system will do, but why it chooses to do it.[7] This, together with the need for explainability and responsibility leads towards systems with an identifiable central decision-making software component and, in this case, the formal verification of this software component can ensure that its decision-making is correct and allows us to analyse the decisions an autonomous system makes against the decisions that a human driver should

take. In complementary work, in aerospace[8], it is formally verified that an autonomous (air) vehicle always follows (selected) "Rules of the Air".

In the past 2-3 years, work has also started on how formal verification methods could be applied to making the problem of verifying HAVs easier. For example, by understanding and formalizing specific desired granular driving behaviours and checking by conventional means that any SUT can be verified to satisfy those granular behaviour requirements, the goal would be to eliminate collisions by design. Two notable contributions have come from TU Munich and MobilEye respectively as first attempts at producing a formal mathematical model for acceptable driving behaviour, using a concept of measuring and determining blame in the case of a collision.[9,10]

These are useful contributions to the process of HAV verification but the work so far is insufficient, not least in that proposed formulation for defining blame-free behaviour as set out in the most recent paper (for example in the presence of a child playing near parked cars) would imply a vehicle speed of just 10-15mph, yet humans can and do drive safely at 20-25mph in the same scenarios. Work is needed to consider how those formulations can discover and capture the more complex processes that humans are using for driving, including the social norms, customs and behaviours that are an essential (locally specific) element of the safe driving in mixed human/HAV environments.

COMPONENTS OF A HIGHLY AUTOMATED VEHICLE CERTIFICATION PROCESS

New Type Approval Process

The current Whole Vehicle Type Approval is well-suited to the present model where those component, system and vehicle specifications can be well-defined, stable, recorded, tested and approved.

But this process is clearly inadequate for HAVs, because:

- Many requirements will be highly specific to the ODD
- Requirements will change on a continuous basis as new vehicles, objects, behaviours, signage are emergent in the ODD over the life of the SUT and HAV
- There will always be a gap to the real-world requirements, necessitating a continuous process of requirements discovery through vehicle testing and/or live vehicle logging
- On discovery of a failure, HAV technology developers will be obligated to provide updated software and models to the vehicle and/or upgrade sensors, compute, communications technology or other AD capabilities and this could be very frequent and/or urgent
- Those updates, in improving performance on certain identified failure modes, may cause unexpected regressions or changes in others

The safety risks from AD operation are very different to those being managed in the current Whole Vehicle Type Approval process and a new testing and certification process is required.

To protect the public, improve road safety over time, assure public trust and ensure our economies and citizens reap the benefits of HAVs ahead of other developed economies:

- A new **HAV Type Approval** process must discover and establish a very high safety standard for the verification and certification of any SUT used on public roads in the UK and, if adopted, across the EU
- The certification of HAV Type Approval, at least initially, must be specific to the requirements of a well-defined ODD and a well-defined DHC; those requirements must represent a complete specification, following extensive discovery
- That high standard must be consistently applied to all HAVs seeking certification for any ODD/DHC pair

- Each certification of HAV Type Approval granted for any ODD/DHC pair must carry the obligation of Conformity of Production, meaning that all subsequent hardware and/or software changes must not reduce the overall safety of the design measured against the then current and most complete ODD/DHC requirements specification; the meaning of overall safety in this context will need to be established, likely as a high threshold pass rate of a statistically significant sample of regression and generative test cases in an ODD/DHC pair in full stack and motion planning simulation environments

- Any request for a HAV to operate outside its ODD/DHC pair must be accompanied by a further certification process for the changed ODD or DHC respectively

Practical Aspects of Validation, Verification and Certification

Scenario Sharing As discussed, discovering scenarios that can inform safe system requirements for HAVs is expensive work spanning large scale, multi-fidelity simulation as well as physical testbed and public road testing. Much like the expensive process of drug discovery, no commercial organization could incur the expense and take the commercial risk unless they were assured some preferential use of the resulting outputs, which in the case of drug molecule development and testing, is achieved through patent protection. For the organizations developing HAV technology, requirements discovery is a similarly huge investment but also a source of competitive advantage.

Clearly a balance needs to be struck between sharing discovered requirements for the public good and that commercial imperative, without retreating to a legally enforceable patenting regime.

We propose a model by which HAV technology developers are encouraged to share the scenarios they discover with the UK certification authority (UKCA). Independent test houses would be commissioned by the UKCA and provided with controlled access to the scenarios for the purposes of evaluating HAV performance for both certification and also on behalf of regulated UK insurers, where required:

- Submitted scenarios are evaluated by UKCA for possible acceptance into an ODD certification test catalog, for example on the basis of probability in the target ODD or on the basis of more than one HAV technology developer executing the scenario and passing the test

- HAV technology developers can elect for submitted and accepted scenarios to be made public, but are not obligated to do so

- Any SUT for any target ODD must be tested in simulation by an independent test house against the full test catalog applicable for the ODD certification (whether publicly visible or not)

- A publicly described test oracle will determine whether a test has passed or failed, based on an overall safety threshold set by the UKCA in which the probability of occurrence of each scenario for the target ODD must be evaluated

- Where a SUT fails a private scenario, abstracted feedback would be provided to the HAV technology developer of that SUT, e.g. SUT failed in interaction with a cyclist.

A process along these lines has the following advantages:

- All HAV technology developers are encouraged to submit scenarios for testing in order to raise the bar for competitors seeking certification in an ODD

- Abstracted feedback from failed tests should encourage HAV technology vendors to generally improve their system safety performance rather than ‘gaming’ a solution to a specific scenario. However, the UKCA must ensure that vendors are not prevented from passing certification by being required to pass highly unlikely scenarios for which they are not provided details

- A market is created for non-competing HAV vendors, e.g. component suppliers or others, to find and submit private scenarios which, once accepted by the approval body have in themselves a value which can be licensed to technology vendors seeking system certification

- Independent test houses would not be subject to the same Freedom of Information (FOI) requests as a government body and could thus protect HAV technology vendors from being forced to disclose exhaustive details about their performance in relation to specific scenarios which could result in the disclosure of valuable trade secrets

Scenario Validation Process The validation of submitted scenario candidates into an ODD test catalog that becomes mandatory is a process which must be developed.

Lessons can likely be drawn from other sectors which have successfully tackled similar challenges of competitive technology development which must evolve standards and operate in a shared common environment.

Cellular wireless telecommunications is perhaps a good example, where the establishment of a cross-industry body, the Global Certification Forum (GCF) was established to define the priority of different work items (in their case dependent on Mobile Network Operators' deployment plans) and how to test conformance to relevant Third Generation Partnership Project (3GPP) wireless standards, such as 3G and LTE, so that the standard is met and there is a strong basis for expecting inter-operability between different networks and network equipment. GCF defines work items to prioritize specific test areas, working groups are drawn from industry participants to review those work items and to seek agreement on the ingredients of conformance test cases, pass criteria, parameters for simulation etc. and to be the final determinant of formal adoption of test cases as part of the mandatory program to be certified as GCF compliant by independent test houses. Obligatory test cases for each work item grow as standards are evolved and field failures are identified. In the example of 3G standards, the mandatory test program as a whole escalated quickly from tens of test cases to thousands of increasingly complex test cases over several quarters. In GCF's case, for any new test case to be adopted, it must be shown to be reproducible and repeatable, which normally means that two separate test and measurement organizations must demonstrate the implementation and execution of the same test. Cellular wireless technology development has different market and technology dynamics (global standards, established test and measurement companies, already a highly competitive market, contained functionality, and not safety critical, etc.), so an exact read across to HAVs will not work, but adaptation of some of these ideas for certifying HAVs could be instructive.

Simulation and Test Tool Sharing HAVs will ultimately operate over a wide range of real-world environments, some of which will be extremely complex with enormous possible state spaces and failure conditions to explore and verify. That leads to the conclusion that simulation must play a lead role in any effective testing procedure, in the generation of test conditions, in the parallelization and/or faster than real-time scaling up needed in the measurement of test coverage and in the defect-seeking capabilities of the randomization testing.

Closed course, physical testbed testing is one form of simulation that has a role to play in any meaningful testing regime. It serves an important purpose in that a real HAV's full stack response is measured with hardware in-the-loop within a full physical environment that has been designed to stress known specific risk aspects of the system and can not only identify system defects against those specific scenarios but can also pinpoint simulation modelling defects and gaps to the real world.

However, the dominant focus of any robust certification process should rest on system verification using high-fidelity software simulation at hyper-scale across a vast number of permutations and combinations. Moreover, this process must leverage tools to explore state space and seek defects, such as fuzzing and directed random testing as well as replaying regression suites of curated scenarios.

Engaging private enterprises in developing and contributing to the development and operation of this certification process is a real consideration, particularly in relation to parties who do not themselves plan to be operators of HAVs in the target ODDs, as may be the case for some HAV technology developers. Even in these cases, there may be a preference for reserving tools and models as trade secrets over sharing their utility across the industry as a whole.

Governments therefore may have an important role to play in enabling a market to exist for the licensing of tools or for their commercial use by practising entities. The objective needs to be to ensure that development expenses required for necessary and valuable independent simulation and tools are capable of being leveraged into meaningful

revenue streams by technology developers. One means of achieving this would be for government to signal and enable a market for such tools to be created, for example through the UKCA or to sponsor a cross-industry unit to step up to the important task, one possible candidate being the UK Government's Meridian initiative.

Non-Deterministic Behaviour Verification Process The behaviour of SUT for HAVs will exhibit non-determinism in the sense that if we repeat what is an identical test execution with what we believe to be an identical opening state, we might still get a different system behaviour. This non-determinism may be a result of e.g. random noise injection, race conditions, or some other aspects of operating system performance.

Therefore, in order to build confidence in system performance, an effective verification program may need to run a single test case multiple times. Where possible, that program must eventually reason about the number of test executions necessary to achieve a defined confidence level in the results, using probabilistic arguments.

Where possible, however, the industry should seek to build tools that offer repeatability and possibly even random stability to ensure that defects can be re-found and corrective actions can be proven to have been effective.

Test Oracle A test oracle is a mechanism for determining whether a system has passed or failed a test and usually is comprised of three capabilities:

- A generator, to provide predicted or expected results for each test
- A comparator, to compare predicted and obtained results
- An evaluator, to determine whether the comparison results are sufficiently close to be a pass

Any of the oracle capabilities may be automated and an automated test oracle will be required to generate, compare and evaluate the performance of the SUT across the test scenario catalog, and perhaps in a fully generative model within the ODD constraints, to ultimately determine if the system performed acceptably given the certification criteria.

The generator should make use of the DHC as extended and the comparator should compare the SUT results against the desired DHC and safety outturns. Evaluation is significantly more complex than simply determining if the SUT was involved in a collision since at one extreme bad driving behaviour doesn't always result in a collision and at the other, a collision is the safest choice for a given set of circumstances.

The specification for the test oracle should be made available to HAV technology developers seeking certification.

Conformity of Production (CoP) Audit Conformity of Production (CoP) is a means of evidencing the ability to produce a series of products that exactly match the specification, performance and marking requirements outlined in the type approval documentation.

In the context of HAVs and a new certification process, HAV technology developers will need to provide evidence to the satisfaction of UKCA that the HAV SUT is representative of all of that Type and that the process of developing and deploying design changes is robust. The form of evidence will need to be carefully considered but could follow the lines of a process review.

The International Automotive Task Force (IATF) together with the International Organization for Standardization (ISO) has developed a standard, IATF16949:2016.[11] This defines the quality management system requirements for the design and development, production, installation and service of automotive-related products.

To achieve IATF certification, an automotive supplier has to work according to automotive core tools, such as:

- Advanced Product Quality Planning – a structured approach to the design and development of products and processes
- Production Part Approval Process – formal release by the customer of a supplier's product and process

- Failure Mode Effect Analysis – risk analysis tool in which a supplier analyses the major risks of not fulfilling the required functions in the current design or process
- Measurement System Analysis – evaluation of the reliability of the measurement systems used by a supplier in its process
- Statistical Process Control – a method of quality control which uses statistical methods to monitor and control a process
- 8D Problem Solving – structural approach to analyze problems, including root causes analysis, containment and corrective actions[12]

Since IATF16949:2016 contains all the key elements for a QMS and is already centered on automotive applications, it is a strong candidate to deliver the framework for the CoP audit compliance of HAVs.

PROMISING NEW RESEARCH AREAS

Swiss Cheese Model Recent promising research at TU Darmstadt has centered on applying a technique known as the Swiss Cheese Model to HAV verification for assessing the probability of collision. In essence, each sensing modality, process, behavioural or environmental variable has 'holes' which could permit a failure and when those holes line up, a collision can occur. One of the key unknowns for assessing the safety of HAVs versus human drivers is a strong understanding of the gap between the probability of critical situations arising in driving scenarios and probability that those critical situations do actually result in a collision. In human driven cars, this difference is a representation of the driving skill and attention of the driver themselves. But on replacing the human with the SUT, those human failure modes (which could be inattention, blind spots etc.) are replaced with new failure modes (which could be detection and classification accuracy, prediction failures etc.). The replacement of one set of cheese slices with another can exhibit quite different failures which demands further research.

Quantifying and measuring these impacts has the potential for us to measure the probability of collision and to compare the two in quantifiable ways and deserves further research.[13]

Extreme Value Theory In another initiative, this time from a research team at Volvo Cars, studies into the use of near-collision measurement as a means of estimating the frequency of actual collisions show good promise. Their approach uses a technique called Extreme Value Theory but more importantly highlights the need for further research into capturing and using near collision data for robust collision rate estimation.[14]

These ideas, and many others, should be reviewed and considered for the on-going development of HAV validation, verification and certification processes.

Remaining Research Questions Key research questions remain, and industry participants can and should work together with leading academics in UK and EU to address them, including:

- What machine learning methods should be applied to directed random testing in coverage driven verification?
- Given the high dimensionality of the test space, what coverage measures are meaningful in generative and ODD digital twin verification?
- Which computer vision methods can we apply to the 3D reconstruction and annotation of digital twin worlds from photogrammetry, Lidar scans and other sensing modalities that mean accurate, up-to-date digital twins are feasible?
- What hardware acceleration beyond GPUs can we design and apply to enable real-time and faster-than-real-time full stack verification of HAVs?

- How can we apply formal software checking to the complex integrated systems required for autonomous driving to ensure that each build achieves its goals without bugs or gaps?
- How do we really apply formal mathematical methods to fully express the DHC, vehicle dynamics and other road user expectations and behaviours to allow us to verify the behavioural safety of HAVs?
- How can we verify HAV systems that comprise of one or more end-to-end neural networks with the requirements to explain failure modes and take corrective actions to improve their performance using human readability and intermediate outputs of modular processes?
- How might we extrapolate randomized testing, including near collisions, into a measure of probability of collision generally?

QUESTIONS FOR POLICY MAKERS

This paper makes a number of suggestions that fall into the realm of policy making, including:

- Could a validation, verification and certification system, such as that outlined in this paper, that incentivizes sharing of scenarios while protecting the value intrinsic to their discovery, improve safety across the industry? Could it be used by an approval body such as UKCA to establish a high standard for UK certification?
- Can the industry agree on a scenario description language that supports coverage-driven verification and is extensible? Is Pegasus a suitable basis for extension to meet this?
- What should the specification of an appropriate simulation environment be and would the government request to tender for delivery of such a tool?
- Could the specification for a test oracle be made available and could this be based on a formal description of 'good driving' in accordance with a DHC?
- Is auditable adherence to the IATF16949:2016 quality assurance process sufficient to satisfy 'Conformity of Production'?

CONCLUSIONS

A number of different methods must be combined for the robust certification of HAVs for deployment in ODDs in the United Kingdom and, by extension, other jurisdictions in Europe.

At the centre of this process is hyper-scale fuzzed scenario-based testing and the use of coverage driven verification methods in digital twins of the ODD and using generative models representative of each ODD. Testing must cover both full stack testing, which will require photo-realistic and sensor-realistic rendering of scenarios and objects, together with accurate sensor modelling and motion planning stack testing, which will require robust beliefs over actor behaviours to test predictive, planning and motion synthesis capabilities. A method for sharing scenarios to a UKCA for industry-wide testing will be required and a means of balancing that sharing for the public good with the need to retain economic leverage over the necessary costs of discovering those requirements will need to be devised. A DHC to include good driving behaviours will be needed and a test oracle will be required to evaluate and publish certification performance.

REFERENCES

[1] The specific conditions under which a given driving automation system or feature thereof is designed to function, including, but not limited to, driving modes is known as the ODD, in accordance with SAE J 3016

[2] "Even if the safety of autonomous vehicles is low—hundreds of failures per 100 million miles, which is akin to human-driven injury and crash rates—demonstrating this would take tens or even hundreds of millions of

miles, depending on the desired precision." Kalra, Nidhi and Susan M. Paddock, Driving to Safety: How Many Miles of Driving Would It Take to Demonstrate Autonomous Vehicle Reliability?. Santa Monica, CA: RAND Corporation, 2016. https://www.rand.org/pubs/research_reports/RR1478.html.

[3] Number of motor insurance claims notified in the United Kingdom (UK) from 2010 to 2016 (in millions), Statista 2018

[4] http://www.pegasus-projekt.info/en/

[5] Huang X., Kwiatkowska M., Wang S., Wu M. (2017) Safety Verification of Deep Neural Networks. In: Majumdar R., Kunčak V. (eds) Computer Aided Verification. CAV 2017. Lecture Notes in Computer Science, vol 10426. Springer, Cham

[6] Monitor/actuator pair architectures can be used in-vehicle to separate the most complex autonomy functions from simpler safety functions. The primary AV functions are performed by the high complexity 'actuator' system, and a paired module (the 'monitor') performs an acceptance test / behavioural validation. The low complexity monitor system should be more straightforward to verify and therefore could, potentially, be verified to ISO 26262 ASIL-D

[7] Fisher, Dennis, Webster: Verifying Autonomous Systems. Communications of the ACM 56(9):84-93, 2013. http://doi.acm.org/10.1145/2494558

[8] Webster, Cameron, Fisher, Jump: Generating Certification Evidence for Autonomous Unmanned Aircraft Using Model Checking and Simulation. J. Aerospace Inf. Sys. 11(5):258-279, 2014. https://doi.org/10.2514/1.I010096

[9] Rizaldi, A., & Althoff, M. (2015, September). Formalising traffic rules for accountability of autonomous vehicles. In Intelligent Transportation Systems (ITSC), 2015 IEEE 18th International Conference on (pp. 1658-1665). IEEE.

[10] Shalev-Shwartz, S., Shammah, S., & Shashua, A. (2017). On a Formal Model of Safe and Scalable Self-driving Cars. arXiv preprint arXiv:1708.06374.

[11] http://www.aiag.org/quality/iatf16949/iatf-16949-2016

[12] "Working in the automotive industry" H. Broekman; D. Ekert; M.I. Kollenhof A.E. Riel; H.C. Theisens; R. Winter, 2017

[13] Winner, H., Wachenfeld, W., & Junietz, P. (2018). Validation and Introduction of Automated Driving. In Automotive Systems Engineering II (pp. 177-196). Springer, Cham

[14] Åsljung, D., Nilsson, J., & Fredriksson, J. (2016). Comparing Collision Threat Measures for Verification of Autonomous Vehicles using Extreme Value Theory. IFAC-PapersOnLine, 49(15), 57-62